

Cell Systems, Volume 12

Supplemental information

**AutoGeneS: Automatic gene selection
using multi-objective optimization
for RNA-seq deconvolution**

Hananeh Alee and Fabian J. Theis

AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution

Hananeh Aliee¹ and Fabian J. Theis^{1,2,3,*}

¹Institute of Computational Biology, Helmholtz Centre, Munich, Bayern, 85764, Germany

²Department of Mathematics, Technische Universität München, Munich, Bayern, 85748, Germany

³Lead Contact

*Correspondence: fabian.theis@helmholtz-muenchen.de (F. J. T.)

Supplementary material

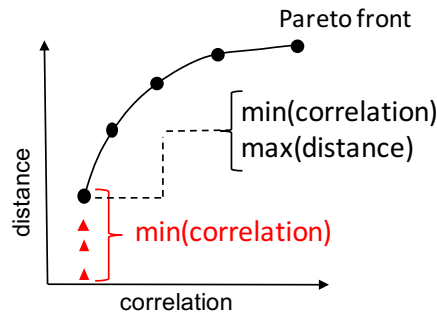


Figure S1: Single- vs multi-objective optimization, Related to Figure 1. Multi-objective optimization finds the solutions that their distance is higher than all the other solutions with the same correlation value.

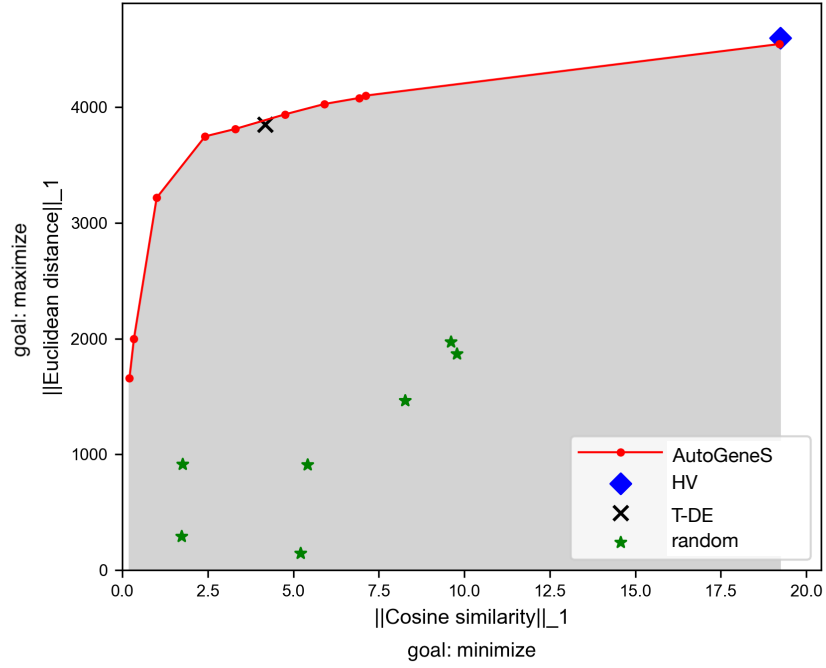


Figure S2: Pareto optimal solutions, Related to Figures 1 and 2. Shown are the objective values, the sum of pair-wise Euclidean distances and correlation coefficients (cosine similarity), for Pareto-optimal solutions using AutoGeneS, and solutions using highly variable (HV) genes, differentially expressed genes using T-test (T-DE), and random sets of genes for human embryonic data from (Chu et al., 2016). Each point refers to a set of signatures. The goal of the optimization was to maximize the Euclidean distances and minimize the correlation coefficients. The space that is explored during the optimization is colored as gray.

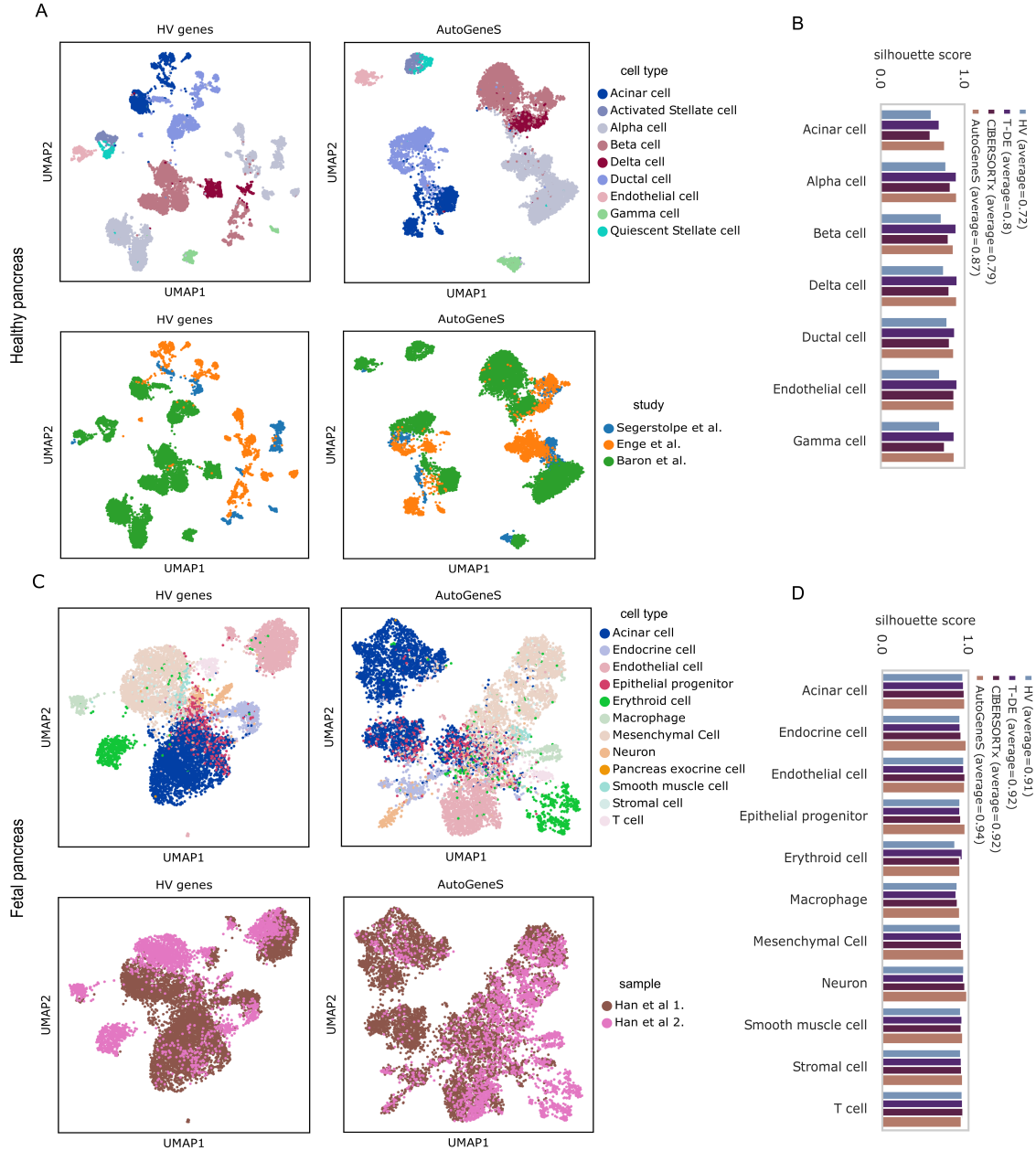


Figure S3: UMAP visualization and Silhouette score for healthy and fetal pancreas, Related to STAR Methods section. The single-cell data were taken from (Baron et al., 2016), (Enge et al., 2017), (Segerstolpe et al., 2016), (Han et al., 2020)). **A** UMAP plots of 10,955 single cells from healthy pancreas (n=3 studies) using HV genes (left) and AutoGeneS (right). Colors highlight cell types and studies. **B** Scaled Silhouette score for each cell type treating each study as a batch. High scaled Silhouette score is desired. **C** UMAP plot of 9,456 single cells from fetal pancreas (n=2 samples). Colors highlight cell types and samples. **D** Scaled Silhouette score for each cell type treating each sample as a batch.

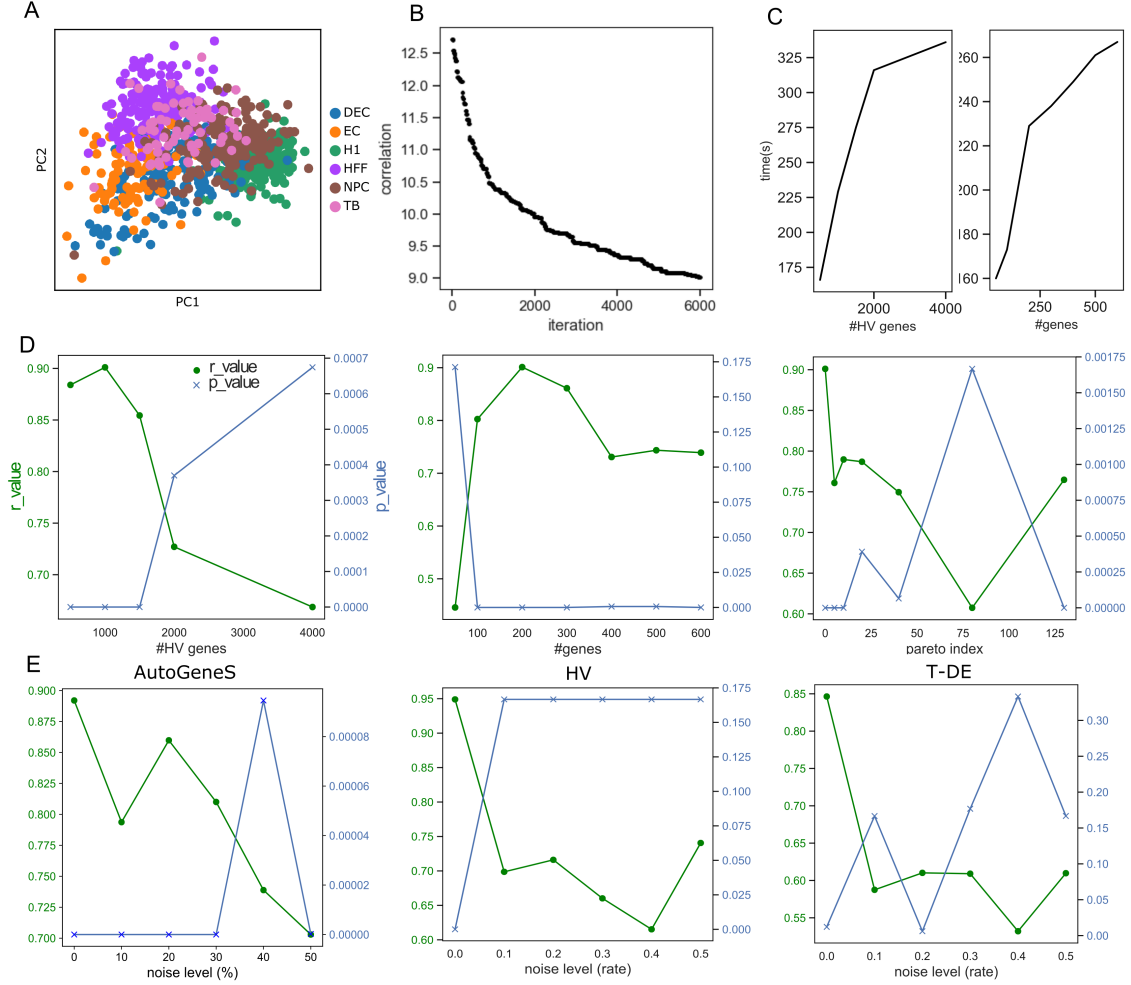


Figure S4: Efficiency analysis of AutoGeneS, Related to Figure 2. To increase the correlation between the cell types and challenge the deconvolution, we removed (i) differentiating ($n=5,379$) and (ii) highly variable genes ($n=5,000$) from the single-cell dataset underlying Figure 2. A total of 8719 genes was left for efficiency analysis. **A** PCA projection of 856 single cells from human embryonic using 1,000 highly variable genes. Shown are PC1 and PC2. **B** Cellular correlation is plotted as a function of optimization run (iteration). Shown is the minimum correlation in Pareto front after each iteration. The optimization ran as long as the minimum correlation converged. **C** Execution time of AutoGeneS as a function of the number of highly variable genes (left) as well as the desired number of genes for deconvolution (right). The number of highly variable genes varies between 500 and 4,000. The number of genes used for deconvolution varies between 50 and 600. **D** r-value and p-value as functions of the number of highly variable genes (left), desired number of genes for deconvolution (middle), and Pareto front index (right). Pareto index=0 refers to the solution with minimum correlation and Pareto index=125 is the solution with the highest distance. **E** r-value and p-value as functions of the noise level. In this analysis, the number of highly variable genes is set to 1,000. Both AutoGeneS and differentially expressed genes based on a t-test selected every time 200 genes from highly variable genes. The noise level was increased from 0% to 50%. For example, noise level=20% means 20 percent of genes from each cell was randomly set to 0.

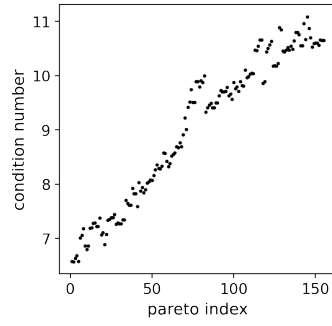


Figure S5: Condition number of Pareto-optimal solutions as a function of Pareto index (noise level=0), Related to Figure S4.

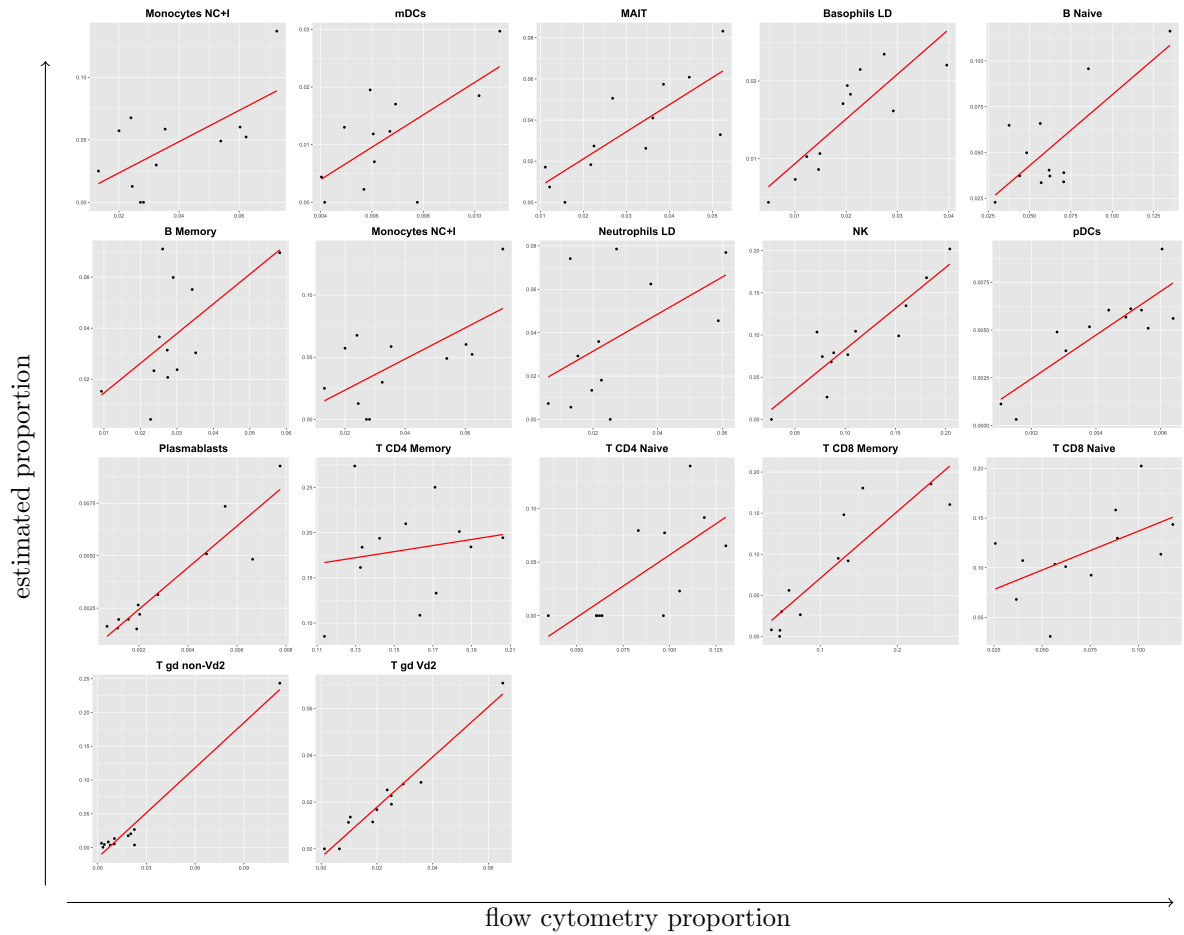


Figure S6: Deconvolution of individual cell types in 12 PBMC samples where sorted RNA-seq is used as the reference data, Related to Figure 3. Shown is the direct comparison between AutoGeneS and flow cytometry for the indicated 17 cell types in PBMC samples. Concordance was determined by linear regression (solid red lines).

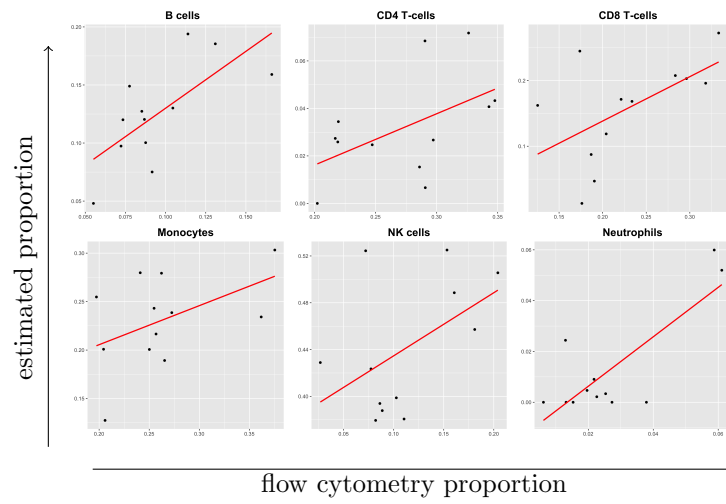


Figure S7: Deconvolution of individual cell types in 12 PBMC samples where single-cell RNA-seq is used as the reference data, Related to Figure 4. Shown is the direct comparison between hierarchical AutoGeneS and flow cytometry for the indicated 6 major cell types in PBMC detected at single-cell resolution. Concordance was determined by linear regression (solid red lines).

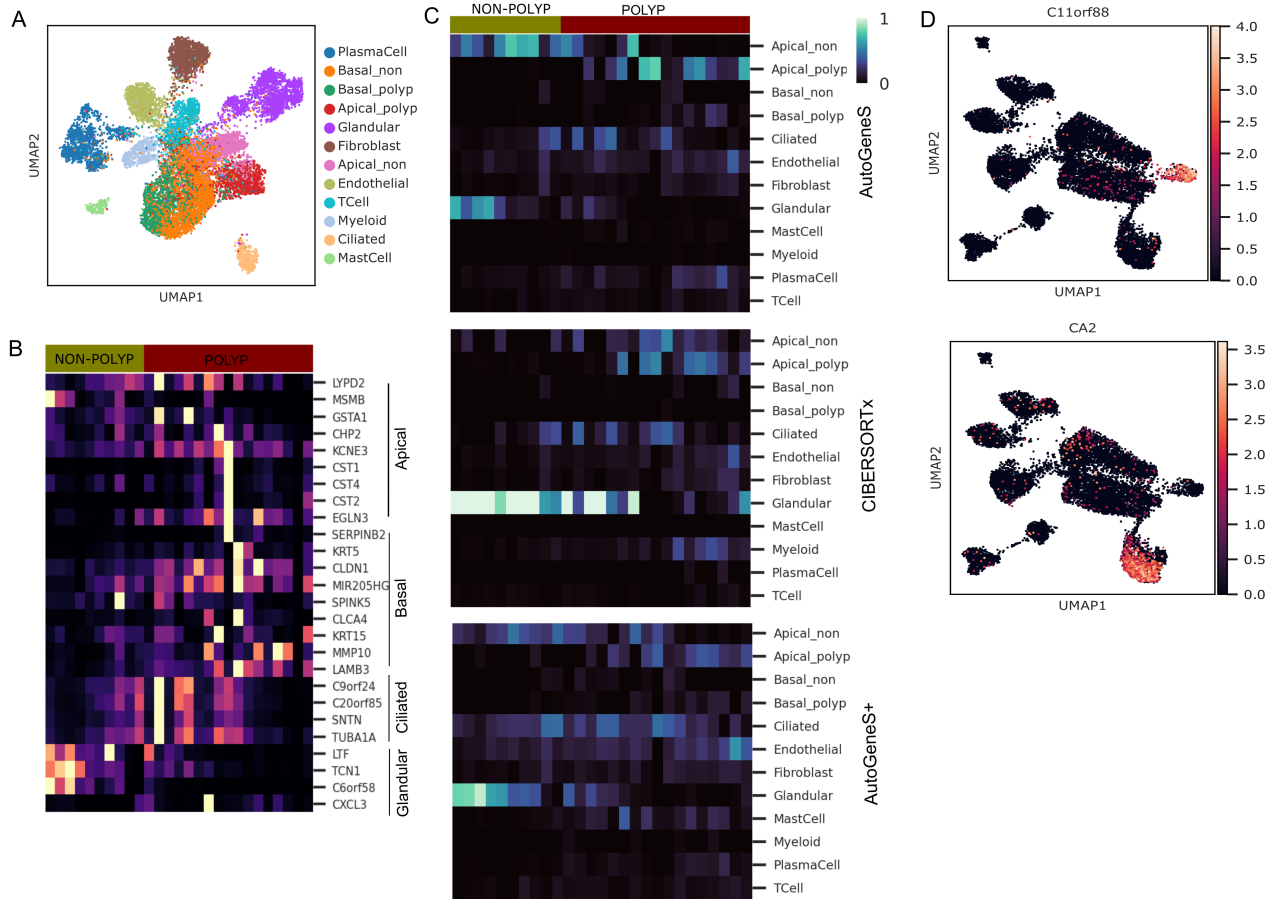


Figure S8: Extended deconvolution results using AutoGeneS and CIBERSORTx, Related to Figure 5. **A** UMAP plot of 14,878 single cells from ethmoid sinus (n=6 non-polyp and n=5 polyp samples) using 400 genes assessed by AutoGeneS. Colors highlight 12 cell types. **B** The expression of signature genes distinguishing apical (secretory), basal, ciliated, and glandular cells across bulk samples. The differentially expressed genes are selected from the 400 discriminative genes using a t-test (see Methods). **C** Proportion of 12 cell types shown in **A** across 27 samples using AutoGeneS and CIBERSORTx. AutoGeneS+ performed the regression using non-normalized counts. **D** UMAP plot colored by the expression of genes C11orf88 and CA2, the biomarkers of ciliated and glandular cells respectively.

crossover and mutation

Ensure:

- 1: *getParents(archive)*: returns two parents from *archive*
- logic_or(parent_1, parent_2)*: returns the logical or of parent_1 and parent_2
- logic_xor(parent_1, parent_2)*: returns the logical xor of parent_1 and parent_2
- nonzero(union)*: returns nonzero indices of list union
- mutation_rate*: returns mutation rate
- ind_size*: returns $|N|$
- nfeatures*: returns $|G|$
- individual*: the individual for mutation

crossover:

Require: nfeatures, archive

- 1: offspring_1, offspring_2 // the new offsprings
- 2: parent_1, parent_2 = getParents(archive)
- 3: union = logic_or(parent_1, parent_2)
- 4: non_zero_id = nonzero(union)
- 5: offspring_1.fill(0)
- 6: offspring_2.fill(0)
- 7: offspring_1[non_zero_id[:nfeatures]] = 1
- 8: offspring_2[non_zero_id[-nfeatures:]] = 1
- 9: **return** offspring_1, offspring_2

mutation:

Require: nfeatures, individual

- 1: n_flip = binomial(ind_size, mutation_rate)
 - 2: n_flip = int((n_flip-n_flip%2)/2)
 - 3: id_1 = np.nonzero(individual)[0]
 - 4: id_0 = np.nonzero(np.invert(individual))[0]
 - 5: ch_1 = self.gen.choice(id_1, n_flip, replace=0)
 - 6: ch_0 = self.gen.choice(id_0, n_flip, replace=0)
 - 7: bm = array([0]*ind_size)
 - 8: bm[ch_1] = 1
 - 9: bm[ch_0] = 1
 - 10: individual = logic_xor(individual, bm)
 - 11: **return** individual
-

Figure S9: Crossover and mutation algorithms, Related to STAR Methods section. The crossover operator combines the bits of the parents to generate new offspring with a fixed number of active bits (the bits with a value of 1). First, the union of the parents is calculated, and the first resulting $|G|$ active bits of that are stored as *offspring_1*. The last $|G|$ active bits are stored as *offspring_2*. The mutation operator makes random changes in the newly generated offsprings. It first selects the number of mutations (*n_flip*) and then randomly flips $n_flip/2$ bits with a value of 0 and $n_flip/2$ bits with a value of 1 in each offspring.

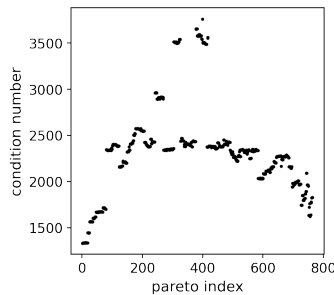


Figure S10: Condition number of Pareto-optimal solutions as a function of Pareto index, Related to Figure 3.

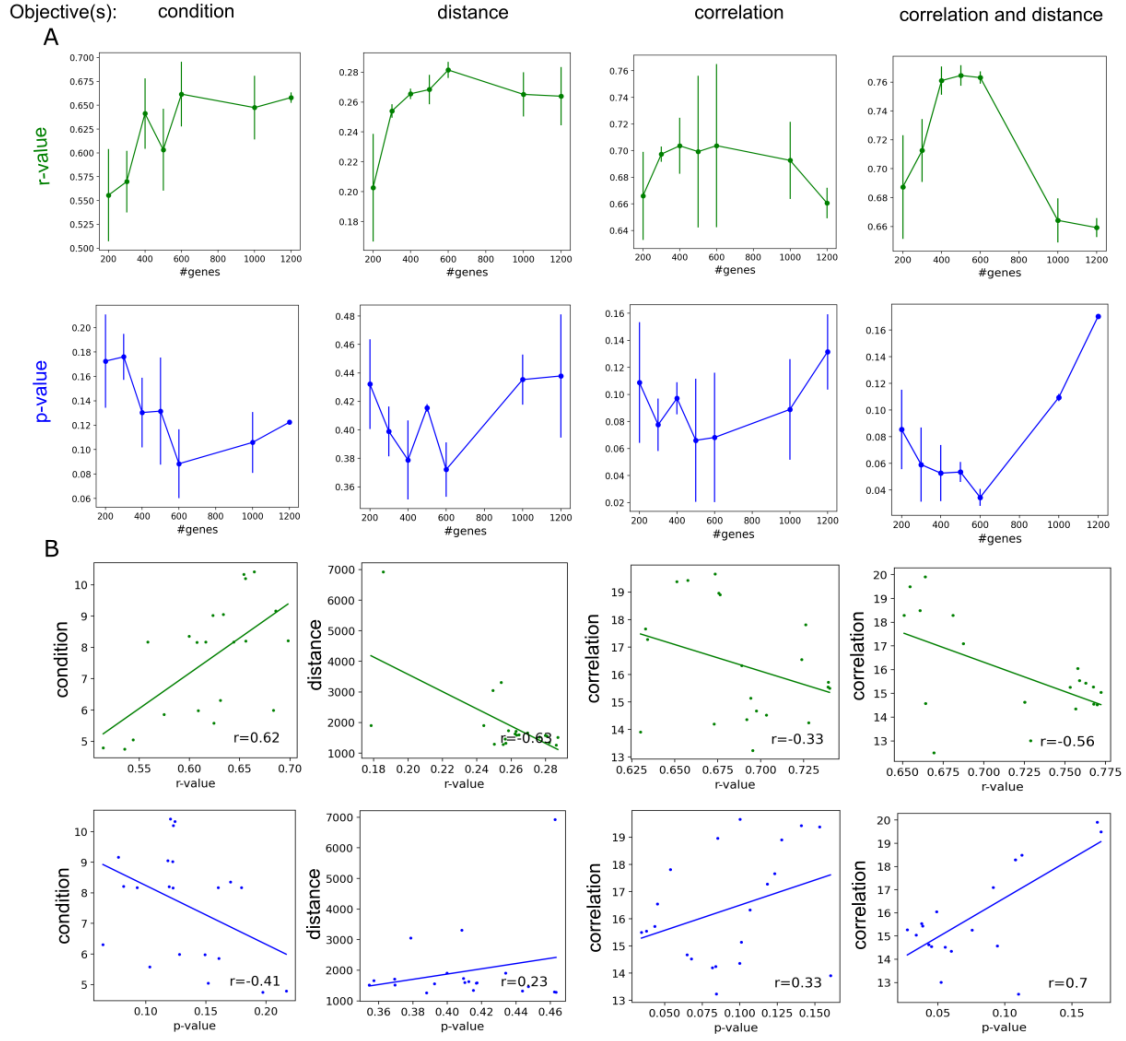


Figure S11: Single vs multi-objective optimization, Related to Figure 3. **A** r- and p-values as functions of the number of genes using single (condition, correlation, or distance) or multiple objectives (correlation and distance) for the dataset underlying Figure 3. The r- and p-values were measured by fitting a linear regression with estimated and real proportions from flow cytometry. Shown are average values and error bars across three runs per #genes. **B** The objective values of the optimization results in **A** as a function of r- and p-values. The significance of the results was assessed by r-value from linear regression (solid green and blue lines).

Supplementary Table 1: Summary of single-cell datasets, Related to STAR Methods section.

dataset	tissue	animal	cells	samples	platform	condition	figure
GSE75748(Chu et al., 2016)	embryo	human	856	1	SMARTer	healthy	2,S4
dataset(Reusch et al., 2021)	blood	human	3,417	1	SeqWell	healthy	4,S7
GSE84133(Baron et al., 2016)	pancreas	human	7,714	4	inDrop	healthy	S3
GSE81547(Enge et al., 2017)	pancreas	human	1,097	1	Smart-seq2	healthy	S3
E-MTAB-5061(Segerstolpe et al., 2016)	pancreas	human	2,544	1	Smart-seq2	healthy	S3
7235471(Han et al., 2020)	pancreas	human	9,769	2	microwell-seq	fetal	S3
dataset(Ordovas-Montanes et al., 2018)	ethmoid sinus	human	14,87	11	Seq-Well	non-polyp,polyp	5,S8